Statistica Sinica **34** (2024), 657-677
doi:https://doi.org/10.5705/ss.202022.0070

INFERENCE FOR PROJECTION-BASED WASSERSTEIN DISTANCES ON FINITE SPACES

Ryo Okano[1] and Masaaki Imaizumi*[1,2]

[1] *The University of Tokyo and* [2] *RIKEN Center for AIP*

Abstract: The Wasserstein distance is the distance between two probability distributions, and has recently become popular in statistics and machine learning, owing to its attractive properties. One important approach to extending this distance is to use low-dimensional projections of the distributions, thus avoiding a high computational cost and the curse of dimensionality in empirical estimation; hare, examples include the sliced Wasserstein and max-sliced Wasserstein distances. Despite their practical success in machine learning tasks, statistical inferences for projection-based Wasserstein distances are limited, owing to the lack of distributional limit results. Thus, for probability distributions supported on finite points, we derive the limit distributions of the empirical versions of the projection-based Wasserstein distances. We examine the general class of distances defined by integrating or maximizing the Wasserstein distances between the low-dimesional projections of two distributions. After deriving the limit distributions, we propose a bootstrap procedure for estimating the quantiles of these distributions from the data. This facilitates asymptotically exact interval estimation and hypothesis testing for these distances. Our theoretical results are based on deriving the distributional limit of empirical Wasserstein distances on finite spaces and the theory of sensitivity analysis in nonlinear programming. Finally, we demonstrate the applicability of our inferential methods using a real-data analysis.

Key words and phrases: Bootstrap, distributional limit, projection-based Wasserstein distances, statistical inference.

1. Introduction

The Wasserstein distance is the distance between two probability distributions, and has attracted considerable interest in the statistics and machine learning literature (Villani (2009); Panaretos and Zemel (2019); Peyré and Cuturi (2019)). This distance is based on the optimal transport problem, and measures the amount of work required to transform one distribution into another. Specifically, given two probability distributions P and Q with finite $p \geq 1$ moments and support in $\mathcal{X} \subset \mathbb{R}^d$, for $d \geq 1$, the p-Wasserstein distance between P and Q is defined as

*Corresponding author.

$$W_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p d\pi(x,y) \right)^{1/p}, \tag{1.1}$$

where $\Pi(P,Q)$ is the set of joint probability distributions with respective marginals that coincide with P and Q, known as couplings. Compared with other measures of distribution closeness, such as the Kullback–Leibler divergence or the total variation distance, the Wasserstein distance has two advantages: (i) it is sensitive to the underlying geometry of the distribution support, and (ii) it does not assume absolute continuity of the distributions. As a result, it has recently become an attractive data analytical tool, particularly in computer vision (Rubner, Tomasi and Guibas (2000); Solomon et al. (2015); Sandler and Lindenbaum (2011)) and natural language processing (Kusner et al. (2015); Zhang et al. (2016)).

Various extensions of the original Wasserstein distance have been proposed to address its shortcomings, mainly its high computational costs and the curse of dimensionality in empirical estimation (Peyré and Cuturi (2019); Weed and Bach (2019)). One important approach is to use low-dimensional projections of the distributions, that is, we compute the Wasserstein distances between low-dimensional projections of the distributions P and Q, instead of comparing P and Q directly. The most representative example of this approach is the sliced Wasserstein distance (Rabin et al. (2011); Bonneel et al. (2015)), which averages the Wasserstein distances between random one-dimensional projections. The sliced Wasserstein distance is an easily computable variant of the Wasserstein distance , because the Wasserstein distance between univariate distributions is easily computed. Another example is the max-sliced Wasserstein distance (Deshpande et al. (2019)), which maximizes the Wasserstein distance between random one-dimensional projections and also has a computational advantage. By considering k-dimensional projections ($1 \leq k \leq d$), the max-sliced Wasserstein distance is generalized to the projection robust Wasserstein (PRW) distance (Paty and Cuturi (2019); Niles-Weed and Rigollet (2019)). The PRW distance captures the difference between two distributions effectively if they differ only in a low-dimensional subspace, and solves the curse of dimensionality in empirical estimation (Niles-Weed and Rigollet (2019); Lin et al. (2021)). Several recent studies have shown that these methods are practical for several machine learning tasks (Lin et al. (2020); Kolouri, Zou and Rohde (2016); Kolouri et al. (2019); Carriere, Cuturi and Oudot (2017); Liutkus et al. (2019)).

The development of inferential tools (e.g., interval estimation or hypothesis testing) for the Wasserstein distance and its extensions is an active research area in statistics. As a basis for inferential procedures, the limit distributions of the empirical versions of these distances have been derived in several specific settings. For example, the limit distributions of the empirical Wasserstein distance have been studied when distributions P and Q are supported in $\mathbb{R}$ (Munk and Czado

(1998); Freitag and Munk (2005); del Barrio et al. (1999); Ramdas, Trillos and Cuturi (2017)) and when they are supported on finite or countable points (Sommerfeld and Munk (2018); Tameling, Sommerfeld and Munk (2019)). The limit distributions of the empirical regularized optimal transport distance on finite spaces, which is an easily computable extension of the Wasserstein distance, have been derived by Bigot, Cazelles and Papadakis (2019) and Klatt, Tameling and Munk (2020). However, for projection-based extensions of the Wasserstein distance, such distributional limit results are not well established, which hinders inferences.

We propose inferential procedures for projection-based Wasserstein distances when the distributions P and Q are supported on finite points. We consider two general classes of distances : the integral projection robust Wasserstein (IPRW) distance, which is defined by integrating the Wasserstein distances between k-dimensional projections of the distributions P, and Q $(1 \leq k \leq d)$ and includes the sliced Wasserstein distance as a special case; and (ii) the PRW distance. As our first contribution, we derive the limit distributions of the empirical IPRW distance and PRW distance with entropic regularization. Second, we show the consistency of the rescaled bootstrap (or m-out-n bootstrap), which enables us to estimate the quantiles of the limit distributions from the data. Consequently, we construct asymptotically exact confidence intervals for these two distances, and obtain new statistics for testing the equality of two distributions. In addition, we extend part of the results to the case where distributions are supported on a countable infinite space with a bounded property. Finally, we apply our inferential methods to a real-data analysis.

As technical contributions, we apply the following two new techniques: (i) a sensitivity analysis, and (ii) entropic regularization. These techniques are necessary to extend the delta method approach (Sommerfeld and Munk (2018)) for the Wasserstein distance to our setting with the IPRW and PRW distances. First, we use sensitivity analysis in nonlinear programming, to investigate how the optimal value of an optimization problem changes when the objective function and the constraints change (Fiacco (1983)). Here we regard the PRW distance as the optimal value of a parametric optimization problem with parameters P and Q, and apply the result of the sensitivity analysis to show its directional differentiability. Second, we add an entropic regularization term (Cuturi (2013)) to the PRW distance, which we refer to as the regularized PRW distance, and then study its distributional limit. The regularization term enables us to specify an optimal transport map and handle its Hadamard differentiability.

This study makes the following contributions to the literature:

- We derive the limit distributions of the empirical versions of the IPRW and regularized PRW distances when the distributions P and Q are supported on finite points.

- We show the consistency of the rescaled bootstrap for the IPRW and regularized PRW distances, enabling us to estimate the quantiles of the limit distributions from the data. This facilitates asymptotically exact interval estimation and hypothesis testing for these distances.

- We show the applicability of our inferential methods using a data- analysis.

1.1. Related work

In addition to the distances we consider, there are several extensions of the Wasserstein distance based on low-dimensional projections, such as the generalized sliced (Kolouri et al. (2019)), tree-sliced (Le et al. (2019)), and distributional sliced (Nguyen et al. (2020)) Wasserstein distances. In addition to the projection-based approaches, Cuturi (2013) proposed the entropic regularization of optimal transport, which can be computed efficiently using an iterative method, called the Sinkhorn algorithm. Goldfeld and Greenewald (2020) proposed the smooth Wasserstein distance, which avoids the curse of dimensionality in its estimation by smoothing out local irregularities in the distributions P and Q using a convolution with a Gaussian kernel.

Statistical inference for the Wasserstein distance and its extensions has been studied in several specific settings, based on their distributional limit results. When the distributions P and Q are supported in $\mathbb{R}$, the Wasserstein distance between them has a closed form, and is described as the L^p norm of the quantile functions of P and Q. Using this fact, researchers have studied the limit distributions of the empirical Wasserstein distances in the univariate case and the validity of the bootstap (Munk and Czado (1998); Freitag and Munk (2005); Del Barrio, Giné and Utzet (2005); Ramdas, Trillos and Cuturi (2017)). Inference for the Wasserstein distance over finite spaces is studied by Sommerfeld and Munk (2018), with their results later extended to the case of countable spaces by Tameling, Sommerfeld and Munk (2019). Inference for the entropic regularized optimal transport distance on finite spaces has been studied by Bigot, Cazelles and Papadakis (2019) and Klatt, Tameling and Munk (2020). In a general setting, del Barrio and Loubes (2019) establish central limit theorems for the empirical Wasserstein distance, and Mena and Weed (2019) establish similar results for the entropic regularized optimal transport distance. However, these results contain unknown centering constants that hinder their use for statistical inference.

To the best of our knowledge, statistical inference for projection-based Wasserstein distances has been considered in only one study. Manole, Balakrishnan and Wasserman (2019) propose confidence intervals with finite-sample validity for the sliced Wasserstein distance, and show their minimax optimality in length. Owing to the closed-form expression of the one-dimensional Wasserstein distance, their inference method is valid without imposing strong assumptions, such as the restriction to finite spaces. However, their approach is not applicable

when the projection dimension is greater than one, in contrast to the approach we adopt here.

1.2. Notation

Let $\|\cdot\|$ and $\langle\cdot\rangle$ denote the Euclidean norm and inner product, respectively. Furthermore, $\mathbb{R}_{>0}$ denotes the positive real numbers, $\mathbb{R}_{\geq 0}$ denotes the nonnegative real numbers and $\otimes$ is the Kronecker product. For any $a, b \in \mathbb{R}$, $a \wedge b$ denotes the minima of a and b. For $1 \leq k \leq d$, the set of $d \times k$ matrices with orthonormal columns is denoted as $S_{d,k} = \{E \in \mathbb{R}^{d\times k} : E^\top E = I_k\}$. Note that when $k = 1$, $S_{d,k}$ coincides with the d-dimensional unit ball, $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. Given a map $T : \mathbb{R}^d \to \mathbb{R}$ and Borel probability measure P supported in $\mathbb{R}^d$, $T_\# P$ denotes the pushforward of P under T, defined by $T_\# P(B) = P(T^{-1}(B))$, for all Borel sets $B \subset \mathbb{R}^d$. For any set $A \subset \mathbb{R}^d$, its diameter is denoted by $\operatorname{diam}(A) = \sup\{\|x - y\| : x, y \in A\}$. In addition, $\mathcal{P}(\mathbb{R}^n)$ denotes the set of all subsets of $\mathbb{R}^n$, $\xrightarrow{d}$ denotes convergence in distribution of the random variables, and $\overset{d}{=}$ denotes the distributional equality of the random variables.

2. Background

2.1. Wasserstein distance

In this study, we restrict the support $\mathcal{X} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ to a finite set of size $N \in \mathbb{N}$. Every probability measure on $\mathcal{X}$ is represented as an element in an $(N-1)$-dimensional simplex $\Delta_N = \{r \in \mathbb{R}^N_{>0} : \sum_{i=1}^N r_i = 1\}$; hence, we do not distinguish between a vector $r \in \Delta_N$ and its corresponding probability distribution. Given a support $\mathcal{X} = \{x_1, \ldots, x_N\}$ and order $p \geq 1$, we define a cost vector $c_p(\mathcal{X}) \in \mathbb{R}^{N^2}$ as $c_p(\mathcal{X})_{(i-1)N+j} = \|x_i - x_j\|^p$, for $1 \leq i, j \leq N$, representing the transport cost from x_i to x_j. The p-Wasserstein distance between two distributions $r, s \in \Delta_N$ on $\mathcal{X} \subset \mathbb{R}^d$ is given by

$$W_p(r, s; \mathcal{X}) = \left\{ \min_{\pi \in \Pi(r,s)} \langle c_p(\mathcal{X}), \pi \rangle \right\}^{1/p}, \tag{2.1}$$

where $\Pi(r, s)$ is the set of vectors of length N^2 that represent the couplings of r and s. Formally, $\Pi(r, s)$ is defined as

$$\Pi(r, s) = \left\{ \pi \in \mathbb{R}^{N^2} : A\pi = (r \ s)^\top \right\}, \tag{2.2}$$

where A is a coefficient matrix:

$$A = \begin{pmatrix} I_{N\times N} \otimes 1_{1\times N} \\ 1_{1\times N} \otimes I_{N\times N} \end{pmatrix} \in \mathbb{R}^{2N\times N^2}.$$

The constraint $A\pi = (r, s)^\top$ ensures that π satisfies the marginal constraints: a matrix $\widetilde{\pi} \in \mathbb{R}^{N\times N}$, generated by π as $\widetilde{\pi}_{i,j} = \pi_{(i-1)N+j}$, satisfies $\sum_{j=1}^{N} \widetilde{\pi}_{i,j} = r_i$ for $1 \le i \le N$, and $\sum_{i=1}^{N} \widetilde{\pi}_{i,j} = s_j$ for $1 \le j \le N$.

2.2. Entropic regularization

Entropic regularization is an extension of the Wasserstein distance (Cuturi (2013)). Given $p \ge 1$, distributions $r, s \in \Delta_N$, and a parameter $\lambda > 0$, we define an entropic regularized optimal transport problem as

$$\min_{\pi\in\Pi(r,s)} \langle c_p(\mathcal{X}), \pi\rangle + \lambda\varphi(\pi), \tag{2.3}$$

where $\varphi : \mathbb{R}^{N^2} \to \mathbb{R}$ is the negative Boltzmann- Shannon entropy, defined as

$$\varphi(\pi) = \begin{cases} \sum_{i=1}^{N^2} \pi_i \log(\pi_i) - \pi_i + 1 & \text{if } \pi \in \mathbb{R}^{N^2}_{\ge 0}, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.4}$$

Here, we set $0\log(0) = 0$. Because the problem (2.3) is a strictly convex optimization problem, it has a unique optimal solution, which we refer to as the regularized optimal transport plan $\pi_{p,\lambda}(r, s; \mathcal{X})$. Using this notion, we can define the p-regularized optimal transport distance (or the p-Sinkhorn divergence) between two distributions $r, s \in \Delta_N$ as

$$W_{p,\lambda}(r, s; \mathcal{X}) = \langle c_p(\mathcal{X}), \pi_{p,\lambda}(r, s; \mathcal{X})\rangle^{1/p}. \tag{2.5}$$

Several computational advantages and statistical properties of the regularized optimal transport distance have been studied (e.g., see Cuturi (2013); Peyré and Cuturi (2019); Klatt, Tameling and Munk (2020); Bigot, Cazelles and Papadakis (2019)).

2.3. Projection-based Wasserstein distances

Here, we extend the Wasserstein distance based on low-dimensional projections of the distributions. Fix $k \le d$ and let $\pi_E : x \in \mathbb{R}^d \mapsto E^\top x$, for $E \in S_{d,k}$. For a distribution P on $\mathbb{R}^d$, the k-dimensional projection of P in $E \in S_{d,k}$ is defined by $P_E = \pi_{E\#}P$. That is, P_E is the distribution of $E^\top X$, for $X \sim P$.

IPRW distance: We study k-dimensional projections of the distributions $r, s \in \Delta_N$ on a finite support $\mathcal{X} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$. The Wasserstein distance between the projections of r and s in a direction $E \in S_{d,k}$ is represented by $W_p(r, s; \mathcal{X}_E)$, where $\mathcal{X}_E = \{E^\top x_1, \ldots, E^\top x_N\} \subset \mathbb{R}^k$. The p-IPRW distance (Lin et al. (2021)) is defined as the integral of the Wasserstein distances over the directions E, that is,

$$\mathrm{IW}_p(r,s) = \left(\int_{S_{d,k}} W_p^p(r,s;\mathcal{X}_E) d\mu(E) \right)^{1/p}, \tag{2.6}$$

where μ is a given measure on $S_{d,k}$. This distance is robust against noise if the distributions have low-dimensional structures, and Lin et al. (2021) shows that the IPRW distance with the uniform measure on $S_{d,k}$ solves the curse of dimensionality in estimation. When the projection dimension is $k = 1$ and μ is the uniform measure of $S_{d,1}$, which coincides with the uniform measure on the d-dimensional unit ball $\mathbb{S}^{d-1}$, the IPRW distance corresponds to the sliced Wasserstein distance (Rabin et al. (2011); Bonneel et al. (2015)). The sliced Wasserstein distance has the advantage of being easy to calculate, because the Wasserstein distance between one-dimensional distributions is easy to compute.

PRW distance: The p-PRW distance (Paty and Cuturi (2019)) is defined as the maximum of the Wasserstein distances between k-dimensional projections of $r, s \in \Delta_N$ over directions $E \in S_{d,k}$, that is,

$$\mathrm{PW}_p(r,s) = \max_{E \in S_{d,k}} W_p(r,s;\mathcal{X}_E). \tag{2.7}$$

When $k = 1$, the PRW distance corresponds to the max-sliced Wasserstein distance (Deshpande et al. (2019)). The PRW distance effectively captures the difference between two distributions r and s if they differ only in a low-dimensional subspace, and Niles-Weed and Rigollet (2019); Lin et al. (2021) show that it solves the curse of dimensionality in estimation.

Here, we introduce an entropic regularization for the PRW distance. With a fixed regularization parameter $\lambda > 0$ and projection direction $E \in S_{d,k}$, we represent the regularized optimal transport distance between the projections of r and s as $W_{p,\lambda}(r,s;\mathcal{X}_E)$. Then, the p-regularized PRW distance is defined by

$$\mathrm{PW}_{p,\lambda}(r,s) = \max_{E \in S_{d,k}} W_{p,\lambda}(r,s;\mathcal{X}_E). \tag{2.8}$$

This method with entropy regularization has the advantage of reducing the computational cost, owing to the smoothing out of the nonsmoothness by the maximization (Lin et al. (2020)).

3. Distributional Limits

We study the distributional limits of the empirical version of the IPRW and regularized PRW distances on a finite space. Specifically, we consider the following setting. For probability distributions $r, s \in \Delta_N$ on $\mathcal{X} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ and sample sizes n and m, let $X_1, \ldots, X_n \sim r, Y_1, \ldots, Y_m \sim s$ be independent and identically distributed (i.i.d.) samples. Then, we define their corresponding empirical distributions $\widehat{r}_n, \widehat{s}_m \in \Delta_N$, the ith elements of which are given as

$$\widehat{r}_{n,i} = \frac{\#\{k : X_k = x_i\}}{n}, \quad \widehat{s}_{m,i} = \frac{\#\{k : Y_k = x_i\}}{m},$$

for $1 \leq i \leq N$. Given the order $p \geq 1$ and a regularization parameter $\lambda > 0$, we derive the distributions to which

$$\sqrt{\frac{nm}{n+m}}\{\mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m) - \mathrm{IW}_p(r, s)\}$$

and

$$\sqrt{\frac{nm}{n+m}}\{\mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m) - \mathrm{PW}_{p,\lambda}(r, s)\}$$

converge in law as $n, m \to \infty$. All proofs are deferred to the Supplementary Material.

3.1. Outline and preparation

We derive distributional limits using the delta method, which is based on the differentiability of the IPRW and regularized PRW distances. Specifically, following that $\sqrt{nm/(n+m)}\{(\widehat{r}_n, \widehat{s}_m) - (r, s)\}$ converges to a Gaussian random vector by the central limit theorem, we can derive distributional limits by applying the delta method with the maps $(r, s) \mapsto \mathrm{IW}_p(r, s)$ and $(r, s) \mapsto \mathrm{PW}_{p,\lambda}(r, s)$. To use the delta method in this setting, we consider directional Hadamard differentiability, which is defined as follows.

Definition 1 (Directional Hadamard differentiability (Römisch (2004); Sommerfeld and Munk (2018))). A function $f : D_f \subset \mathbb{R}^d \to \mathbb{R}$ is directionally Hadamard differentiable at $u \in D_f$ tangentially to $D_0 \subset \mathbb{R}^d$ if there exists a map $f'_u : D_0 \to \mathbb{R}$, such that

$$\lim_{n\to\infty} \frac{f(u + t_n h_n) - f(u)}{t_n} = f'_u(h), \tag{3.1}$$

for any $h \in D_0$ and arbitrary sequences $\{t_n\} \subset \mathbb{R}$ and $\{h_n\} \subset \mathbb{R}^d$, such that $t_n \searrow 0$, $h_n \to h$, and $u + t_n h_n \in D_f$, for all large $n \in \mathbb{N}$. We refer to f'_u as the directional Hadamard derivative.

In contrast to the usual (nondirectional) Hadamard differentiability (e.g., van der Vaart (2000)), directional Hadamard differentiability does not require the derivative to be linear, but allows for the delta method.

Theorem 1 (Delta method with a directionally Hadamard differentiable map: Theorem 1 in Römisch (2004)). Let $f : D_f \subset \mathbb{R}^d \to \mathbb{R}$ be directionally Hadamard differentiable at $u \in D_f$ tangentially to $D_0 \subset \mathbb{R}^d$ with the derivative $f'_u : D_0 \to \mathbb{R}$. Let T_n be $\mathbb{R}^d$-valued random variables, so that $\rho_n(T_n - u) \xrightarrow{d} T$ for a sequence of numbers $\rho_n \to \infty$, and a random variable T taking its values in D_0. Then, $\rho_n(f(T_n) - f(u)) \xrightarrow{d} f'_u(T)$.

Our approach based on the directional Hadamard derivative is important for dealing with the projection-based Wasserstein distances. These distances are not differentiable in the sense of (nondirectional) Hadamard differentiation, but do have a directional Hadamard derivative, which makes it possible to apply the delta method.

3.2. Distributional limit for IPRW distance

As our first main result, we derive the distributional limit of the empirical IPRW distance, $\mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m)$. To this end, we first show the directional Hadamard differentiability of the map $(r,s) \mapsto \mathrm{IW}_p^p(r,s)$ and derive its derivative. In preparation, we define the sets of dual solutions for the optimization problem in (2.1). Following Sommerfeld and Munk (2018), given two distributions $r, s \in \Delta_N$ and a ground space $\mathcal{X} = \{x_1, \ldots, x_N\}$, we define the following sets:

$$\Phi_p^*(\mathcal{X}) = \{u \in \mathbb{R}^N : u_i - u_j \le \|x_i - x_j\|^p, 1 \le i, j \le N\}, \tag{3.2}$$

$$\Phi_p^*(r, s; \mathcal{X}) = \{(u, v) \in \mathbb{R}^N \times \mathbb{R}^N : \langle u, r\rangle + \langle v, s\rangle = W_p^p(r, s; \mathcal{X}), \; u_i + v_j \le \|x_i - x_j\|^p, 1 \le i, j \le N\}. \tag{3.3}$$

These sets play a role in describing the limit distributions. In addition, we define the set of directions in which limits are taken as $\Omega_N = \{h \in \mathbb{R}^N : \sum_{i=1}^N h_i = 0\}$. Then, we achieve the following result on differentiability.

Proposition 1 (Directional Hadamard differentiability of IW_p^p). The map $\mathrm{IW}_p^p : \Delta_N \times \Delta_N \to \mathbb{R}, (r,s) \mapsto \mathrm{IW}_p^p(r,s)$ is directional Hadamard differentiable at all $(r,s) \in \Delta_N \times \Delta_N$ tangentially to $\Omega_N \times \Omega_N$, with derivative

$$(h_1, h_2) \mapsto \int_{S_{d,k}} \max_{(u,v) \in \Phi_p^*(r,s;\mathcal{X}_E)} -(\langle u, h_1\rangle + \langle v, h_2\rangle) d\mu(E). \tag{3.4}$$

We state our main result on the limit distribution of the empirical IPRW distance. This derivation is based on the differentiability in Proposition 1 and the delta method in Theorem 1. For $r \in \Delta_N$, we define

$$\Sigma(r) = \begin{pmatrix} r_1(1-r_1) & -r_1 r_2 & \cdots & \text{`} - r_1 r_N \\ -r_2 r_1 & r_2(1-r_2) & \cdots & -r_2 r_N \\ \vdots & \vdots & \ddots & \vdots \\ -r_N r_1 & -r_N r_2 & \cdots & r_N(1-r_N) \end{pmatrix}. \tag{3.5}$$

Then, we obtain the following result.

Theorem 2 (Distributional limits of $\mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m)$). Let $r, s \in \Delta_N$ be two probability distributions supported on $\mathcal{X} \subset \mathbb{R}^d$, $X_1, \ldots, X_n \sim r, Y_1, \ldots, Y_m \sim s$ be i.i.d. n and m samples, respectively, and $\widehat{r}_n, \widehat{s}_m$ be the corresponding empirical distributions. Let $G \sim N(0, \Sigma(r))$ and $H \sim N(0, \Sigma(s))$ be independent Gaussian

random vectors. Then, we have the following

i. If $r = s$, $n \wedge m \to \infty$, and $m/(n+m) \to \delta \in (0,1)$, then we have

$$\left(\frac{nm}{n+m}\right)^{1/2p} \mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m) \xrightarrow{d} \left(\int_{S_{d,k}} \max_{u \in \Phi_p^*(\mathcal{X}_E)} \langle G, u\rangle d\mu(E)\right)^{1/p},$$

where $\Phi_p^*(\mathcal{X}_E)$ is given by (3.2).

ii. If $r \neq s$, $n \wedge m \to \infty$, and $m/(n+m) \to \delta \in (0,1)$, then we have

$$\sqrt{\frac{nm}{n+m}}\{\mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m) - \mathrm{IW}_p(r,s)\}$$
$$\xrightarrow{d} \frac{1}{p}\,\mathrm{IW}_p^{1-p}(r,s) \int_{S_{d,k}} \max_{(u,v) \in \Phi_p^*(r,s;\mathcal{X}_E)} \sqrt{\delta}\langle G, u\rangle + \sqrt{1-\delta}\langle H, v\rangle d\mu(E),$$

where $\Phi_p^*(r, s; \mathcal{X}_E)$ is given by (3.3).

The scaling rate in Theorem 2 is independent of the dimension of the underlying space $\mathcal{X}$, which is the same as those of other extensions of the Wasserstein distance on finite spaces (Sommerfeld and Munk (2018); Klatt, Tameling and Munk (2020); Bigot, Cazelles and Papadakis (2019)). Moreover, for $p > 1$, the scaling rate for $r = s$ (i.e., $n^{-1/2p}$) is slower than that for $r \neq s$ (i.e., $n^{-1/2}$), implying that $\mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m)$ converges more slowly under $r = s$ for $p > 1$. Note that the pth power $\mathrm{IW}_p^p(\widehat{r}_n, \widehat{s}_m)$ has the same scaling rate $n^{-1/2}$, regardless of whether $r = s$ or $r \neq s$.

Although this result focuses on finite spaces, in Section S3 of the Supplementary Material, we derive the distributional limits of IPRW distances on countable infinite spaces with a bounded property.

3.3. Distributional limit for regularized PRW distance

As our second main result, we derive the distributional limit of the empirical regularized PRW distance, $\mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m)$. To study the PRW distance, we need to introduce entropic regularization to add smoothness to the Wasserstein distance. For the regularization of the Wasserstein distance on finite spaces, refer to Klatt, Tameling and Munk (2020).

We derive a distributional limit by showing the directional Hadamard differentiability of the regularized PRW distance, and applying the delta method. Our proof relies on the following results of a sensitivity analysis in nonlinear programming (Fiacco (1983)):

Consider the following general optimization problem with the parameter $u \in U$ in the objective function:

$$\max_{x \in \mathbb{R}^n} f(x,u), \quad \text{subject to} \quad x \in S.$$

Here, $f : \mathbb{R}^n \times U \to \mathbb{R}$ is continuous, and $\nabla_u f$ exists and is continuous on $\mathbb{R}^n \times U$. Moreover, the feasible region $S \subset \mathbb{R}^n$ is a compact set, and the parameter set $U \subset \mathbb{R}^p$ is open and bounded. We define the optimal value function $\phi : U \to \mathbb{R}$ and the optimal set mapping $\Phi : U \to \mathcal{P}(\mathbb{R}^n)$ as $\phi(u) = \max\{f(x,u) : x \in S\}$ and $\Phi(u) = \{x \in S : \phi(u) = f(x,u)\}$, respectively. Then, we have the following result.

Theorem 3 (Theorem 2.3.1 in Fiacco (1983)). For all $u \in U$ and in any direction $h \in \mathbb{R}^p$, the optimal value function ϕ is directionally differentiable in the sense of Gâteaux; that is, the limit (3.1) exists for a fixed h and not a sequence $h_n \to h$. In addition, the derivative is given by

$$h \mapsto \max_{x \in \Phi(u)} \langle \nabla_u f(x,u), h \rangle.$$

We employ this result to demonstrate the directional Hadamard differentiability of the regularized PRW distance.

For a technical reason, we reformulate the regularized optimal transport problem (2.3). The transport condition in (2.2) can be stated in terms of $2N-1$ equality constraints, instead of $2N$, which allows for linearly independent constraints. Following Klatt, Tameling and Munk (2020), we denote by $A_\star$ and $s_\star$ the deletions of the last row of a matrix A in (2.2) and the last entry of a vector $s \in \Delta_N$, respectively. We denote the set of such $s_\star$ as $(\Delta_N)_\star$. Using the constraint $\sum_{i=1}^N s_i = 1$, we identify the vector $s \in \Delta_N$ with $s_\star \in (\Delta_N)_\star$. To apply Theorem 3 to the regularized PRW distance, we show the continuous differentiability of the regularized optimal transport plan with projection in the following lemma.

Lemma 1. *Let $p \geq 2$ and $\lambda > 0$. The map $(r, s_\star, E) \mapsto \pi_{p,\lambda}(r, s_\star; \mathcal{X}_E)$ is continuously differentiable on $\Delta_N \times (\Delta_N)_\star \times \mathbb{R}^{dk}$. In addition, the matrix of partial derivatives with respect to $(r, s_\star)$ at $(r_0, (s_0)_\star, E_0)$ is given by*

$$\nabla_{(r,s_\star)} \pi_{p,\lambda}(r_0, s_{0\star}; \mathcal{X}_{E_0}) = DA_\star^\top (A_\star D A_\star^\top)^{-1} \in \mathbb{R}^{N^2 \times (2N-1)},$$

where $D \in \mathbb{R}^{N^2 \times N^2}$ is a diagonal matrix in which the (j,j)-entry is the jth element of $\pi_{p,\lambda}(r_0, s_{0\star}; \mathcal{X}_{E_0})$.

Now, we show the directionally Hadamard differentiability of the regularized PRW distance. Given $(r, s_\star) \in \Delta_N \times (\Delta_N)_\star$, we define $\Psi_p^*(r, s_\star)$ as the set of directions that maximizes the regularized optimal transport distance between the projections of r and s, that is,

$$\Psi_p^*(r, s_\star) = \{E \in S_{d,k} : W_{p,\lambda}(r, s; \mathcal{X}_E) = \mathrm{PW}_{p,\lambda}(r, s)\}.$$

We denote by $h_\star$ the deletion of the last entry of a vector $h \in \Omega_N$, and the set of such $h_\star$ as $(\Omega_N)_\star$.

Proposition 2. Let $p \geq 2$ and $\lambda > 0$. The map $(r, s_\star) \mapsto \mathrm{PW}_{p,\lambda}(r, s_\star)$ is directionally Hadamard differentiable at all $(r, s_\star) \in \Delta_N \times (\Delta_N)_\star$, tangentially to $\Omega_N \times (\Omega_N)_\star$,with the following derivative:

$$(h_1, h_{2\star}) \mapsto \max_{E \in \Psi_p^*(r,s_\star)} \langle \gamma^\top D A_\star^\top (A_\star D A_\star^\top)^{-1}, (h_1, h_{2\star}) \rangle, \tag{3.6}$$

where

$$\gamma = \frac{1}{p} \langle c_p(\mathcal{X}_E), \pi_{p,\lambda}(r, s_\star, \mathcal{X}_E) \rangle^{1/p-1} c_p(\mathcal{X}_E) \in \mathbb{R}^{N^2}, \tag{3.7}$$

and $D \in \mathbb{R}^{N^2 \times N^2}$ is a diagonal matrix in which the (j, j)-entry is the jth element of $\pi_{p,\lambda}(r, s_\star, \mathcal{X}_E)$, for $j = 1, \ldots, N^2$.

The next theorem states our main result on the limit distribution of the empirical regularized PRW distance.

Theorem 4 (Distributional limit of $\mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m)$). Let $p \geq 2$ and $\lambda > 0$. Under the assumptions of Theorem 2, as $n \wedge m \to \infty$ and $m/(n+m) \to \delta \in (0, 1)$, we have

$$\begin{aligned} &\sqrt{\frac{nm}{n+m}} \{\mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m) - \mathrm{PW}_{p,\lambda}(r, s)\} \\ &\xrightarrow{d} \max_{E \in \Psi_p^*(r,s_\star)} \langle \gamma^\top D A_\star^\top (A_\star D A_\star^\top)^{-1}, (\sqrt{\delta} G, \sqrt{1-\delta} H_\star) \rangle, \end{aligned}$$

where $\gamma \in \mathbb{R}^{N^2}$ and $D \in \mathbb{R}^{N^2 \times N^2}$ are defined in Proposition 2, and $H_\star$ denotes the deletion of the last entry of a random vector $H \sim N(0, \Sigma(s))$.

4. Bootstrap

We approximate the derived limit distributions using a bootstrap procedure. Let $r, s \in \Delta_N$ and $X_1, \ldots, X_n \sim r, Y_1, \ldots, Y_m \sim s$ be i.i.d. samples with empirical distributions $\widehat{r}_n$ and $\widehat{s}_m$, respectively. Furthermore, let $\widehat{r}_\ell^*$ and $\widehat{s}_\ell^*$ be the empirical bootstrap distributions defined by the i.i.d. bootstrap samples $X_1^*, \ldots, X_\ell^* \sim \widehat{r}_n$ and $Y_1^*, \ldots, Y_\ell^* \sim \widehat{s}_m$, respectively.

The functionals IW_p and $\mathrm{PW}_{p,\lambda}$ are only directionally Hadamard differentiable, that is, they have nonlinear derivatives with respect to (h_1, h_2). As mentioned by Dümbgen (1993) and Sommerfeld and Munk (2018), the naive n-out-n bootstrap is inconsistent for such functionals with a nonlinear Hadamard derivative, but that re-sampling fewer than n observations leads to a consistent bootstrap (the rescaled or m-out-n bootstrap). Therefore we obtain the following results on the bootstrap for the IPRW and regularized PRW distances. In the following, $\mathrm{BL}_1(\mathbb{R})$ denotes the set of all bounded functions on $\mathbb{R}$ with a Lipschitz constant of at most one, and $\xrightarrow{*}$ denotes convergence in outer probability (van der Vaart (2000), Sec. 18.2).

Proposition 3. Let $p \geq 1$. We assume that $\ell \to \infty, \ell/n \to 0$, and $\ell/m \to 0$ as $n, m \to \infty$. Then, the plug-in bootstrap with $\widehat{r}_\ell^*$ and $\widehat{s}_\ell^*$ for the IPRW distance is consistent:

i. If $r = s$, $n \wedge m \to \infty$, and $m/(n+m) \to \delta \in (0,1)$, we have

$$\sup_{h \in \mathrm{BL}_1(\mathbb{R})} \left| \mathbb{E}\left[h\left(\left(\frac{\ell}{2}\right)^{1/2p} \mathrm{IW}_p(\widehat{r}_\ell^*, \widehat{s}_\ell^*) \right) \middle| X_1, \ldots, X_n, Y_1, \ldots, Y_m \right] - \mathbb{E}\left[h\left(\left(\frac{nm}{n+m}\right)^{1/2p} \mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m) \right) \right] \right| \xrightarrow{*} 0.$$

ii. If $r \neq s$, $n \wedge m \to \infty$, and $m/(n+m) \to \delta \in (0,1)$, we have

$$\sup_{h \in \mathrm{BL}_1(\mathbb{R})} \left| \mathbb{E}\left[h\left(\sqrt{\frac{\ell}{2}} \{ \mathrm{IW}_p(\widehat{r}_\ell^*, \widehat{s}_\ell^*) - \mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m) \} \right) \middle| X_1, \ldots, X_n, Y_1, \ldots, Y_m \right] - \mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}} \{ \mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m) - \mathrm{IW}_p(r,s) \} \right) \right] \right| \xrightarrow{*} 0.$$

Proposition 4. Let $p \geq 2$ and $\lambda > 0$. We assume that $\ell \to \infty, \ell/n \to 0$, and $\ell/m \to 0$ as $n, m \to \infty$. Then, the plug-in bootstrap with $\widehat{r}_\ell^*$ and $\widehat{s}_\ell^*$ for the regularized PRW distance is consistent. That is, as $n \wedge m \to \infty$ and $m/(n+m) \to \delta \in (0,1)$, we have

$$\sup_{h \in \mathrm{BL}_1(\mathbb{R})} \left| \mathbb{E}\left[h\left(\sqrt{\frac{\ell}{2}} \{ \mathrm{PW}_{p,\lambda}(\widehat{r}_\ell^*, \widehat{s}_\ell^*) - \mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m) \} \right) \middle| X_1, \ldots, X_n, Y_1, \ldots, Y_m \right] - \mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}} \{ \mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m) - \mathrm{PW}_{p,\lambda}(r,s) \} \right) \right] \right| \xrightarrow{*} 0.$$

In practice, the performance of our bootstrap procedure depends on the choice of the replacement number ℓ. In the Supplementary Material, we discuss how the choice of ℓ affects the finite-sample performance of the bootstrap.

5. Applications

5.1. Two-sample testing with sliced Wasserstein distance

Let $r, s \in \Delta_N$ and take $X_1, \ldots, X_n \sim r$, $Y_1, \ldots, Y_m \sim s$ as i.i.d. samples. The nonparametric two-sample test determines whether the sampling distributions r, s are equal, based on samples. This is described as

$$H_0 : r = s \quad \text{vs.} \quad H_1 : r \neq s.$$

Table 1. Rejection rates of the proposed test. The significance level is 0.05.

	$r = s$	$r \neq s$
$\ell = n^{4/5}$	0.001	1.000
$\ell = n^{2/3}$	0.016	1.000
$\ell = n^{1/2}$	0.037	1.000

We propose a test using the sliced Wasserstein distance, that is, the IPRW distance with a one-dimensional projection and a uniform measure. Specifically, we denote $\mathrm{SW}_{m,n} = \sqrt{mn/(m+n)}\mathrm{IW}_p(\widehat{r}_n, \widehat{s}_m)$ and propose the test

$$\mathrm{SW}_{m,n} > c_\alpha \Rightarrow \text{ reject } H_0,$$

where c_α is a critical value chosen based on the given level of $\alpha \in (0,1)$. The two-sample test based on the Wasserstein distance was performed by Ramdas, Trillos and Cuturi (2017). They designed univariate test statistics using the Wasserstein distance, and analyzed their limit distribution. However, their approach is available only for $d = 1$, because it does not extend to higher dimensions. Our proposed test is not restricted to a one-dimensional setting, and can be applied to large-scale data sets because of the low computational complexity of the sliced Wasserstein distance.

We use the bootstrap procedure to choose an appropriate critical value from the data. Let $\widehat{r}^*_\ell$ and $\widehat{s}^*_\ell$ be the empirical bootstrap distributions obtained from the bootstrap samples $X^*_1, \ldots, X^*_\ell \sim \widehat{r}_n$ and $Y^*_1, \ldots, Y^*_\ell \sim \widehat{s}_m$, respectively. We define the bootstrap version of the test statistics as $\mathrm{SW}^*_{m,n} = \sqrt{\ell/2}\ \mathrm{IW}_p(\widehat{r}^*_\ell, \widehat{s}^*_\ell)$, and denote by $\widehat{c}_\alpha$ the $(1-\alpha)$quantile of $\mathrm{SW}^*_{m,n}$. Note that $\widehat{c}_\alpha$ can be computed numerically. Then, the validity of the rescaled bootstrap for the IPRW distance (Proposition 3) implies that, under $\ell \to \infty, \ell/n \to 0$, and $\ell/m \to 0$ as $n, m \to \infty$, the test

$$\mathrm{SW}_{m,n} > \widehat{c}_\alpha \Rightarrow \text{ reject } H_0$$

has asymptotic level α. Specifically, $\limsup_{m,n\to\infty} P(\mathrm{SW}_{m,n} > \widehat{c}_\alpha) \leq \alpha$.

Here, we demonstrate the finite- sample performance of this test. We set the finite ground space $\mathcal{X}$ to be an equidistant two-dimensional 7×7 grid on $[0,1] \times [0,1]$. For the case $r = s$, we generate a distribution $r \sim \mathrm{Dir}(\mathbf{1})$ and set $s = r$, and for the case $r \neq s$, we generate two distributions $r, s \sim \mathrm{Dir}(\mathbf{1})$ independently. We set the sample size as $n = m = 1000$, and vary the replacement number as $\ell \in \{n^{4/5}, n^{2/3}, n^{1/2}\}$. We set the significance level to be $\alpha = 0.05$, and run 1,000 Monte Carlo iterations.

Table 1 shows the rejection rates of the proposed test in each case. For the case $r = s$, the rejection rates should be under the significance level of $\alpha = 0.05$, and this is true for all $\ell \in \{n^{4/5}, n^{2/3}, n^{1/2}\}$. For the case $r \neq s$, the power of the test is 1.000, which is satisfactory.

Figure 1. Data sets of images. The first, second, and third columns show the data sets 1, 2, and 3, respectively.

Table 2. Two-sample testing for the color distributions of images.

Dataset	Statistic	p-value		
		$\ell = n^{4/5}$	$\ell = n^{2/3}$	$\ell = n^{1/2}$
1	15.55	< 0.001	< 0.001	< 0.001
2	9.07	< 0.001	< 0.001	< 0.001
3	0.25	0.446	0.372	0.352

We now apply the proposed test to examine the equality of color distributions in images. Given two different images, the aim is to investigate whether they have significantly different color distributions. Figure 1 shows the data sets of the images used. Each image has $768 \times 576 = 442{,}368$ pixels, and was obtained from a publicly available data set at `http://tabby.vision.mcgill.ca/html/welcome.html`. We transform each image into a color histogram in the RGB color space with grid size $16^3 = 4{,}086$. In data set 1 (the first column in Figure 1), the two images are expected to have different color distributions. In data set 2 (the second column in Figure 1), the two images are expected to have different, but similar color distributions. In data set 3 (the third row in Figure 1), the second image is obtained by flipping the first image around the vertical axis; thus, they have the same color histograms. In each data set, we randomly select $n = 10{,}000$ pixels from each image and construct the empirical color distributions $\widehat{r}_n$ and $\widehat{s}_n$. Then, we calculate the test statistic $\mathrm{SW}_{n,n}$ and the p-values based on $B = 500$ bootstraps with replacement $\ell \in \{n^{4/5}, n^{2/3}, n^{1/2}\}$. Table 2 shows the results.

For data set 1, the proposed test with every replacement ℓ suggests a strong rejection of the null hypothesis. For the data set 2, we also see a strong rejection of the null hypothesis, but the test statistic (9.07) is smaller than that for data set 1 (15.55). For data set 3, the proposed test with any replacement ℓ does not

report a small p-value, which means there is no strong evidence to reject the null hypothesis.

5.2. Interval estimation for regularized PRW distance

Given a level $\alpha \in (0,1)$ and i.i.d. samples $X_1, \ldots, X_n \sim r, Y_1, \ldots, Y_m \sim s$, we construct an asymptotic confidence interval C_{nm} for the regularized PRW distance $\mathrm{PW}_{p,\lambda}(r,s)$, such that

$$\liminf_{n,m\to\infty} P(\mathrm{PW}_{p,\lambda}(r,s) \in C_{mn}) \geq 1-\alpha.$$

The previous distributional results allow us to construct C_{nm}. Although we focus on the regularized PRW distance, we can construct such an interval for the IPRW distance under $r \neq s$ in the same manner.

Let $\widehat{r}_\ell^*$ and $\widehat{s}_\ell^*$ be the empirical bootstrap distributions obtained from the bootstrap samples $X_1^*, \ldots, X_\ell^* \sim \widehat{r}_n$ and $Y_1^*, \ldots, Y_\ell^* \sim \widehat{s}_m$, respectively. We denote the $\alpha/2$ and $(1-\alpha/2)$ quantiles of $\mathrm{PW}_{p,\lambda}(\widehat{r}_\ell^*, \widehat{s}_\ell^*)$ as $q_{\alpha/2}$ and $q_{1-\alpha/2}$, respectively, and define

$$C_{nm} = \left[\mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m) - \sqrt{\frac{n+m}{nm}} q_{1-\alpha/2}, \mathrm{PW}_{p,\lambda}(\widehat{r}_n, \widehat{s}_m) - \sqrt{\frac{n+m}{nm}} q_{\alpha/2}\right].$$

Then, the validity of the rescaled bootstrap for the regularized PRW distance (Proposition 4) implies that under $\ell \to \infty, \ell/n \to 0, \ell/m \to 0$ as $n, m \to \infty$, and $m/(n+m) \to \delta \in (0,1)$, C_{nm} is an asymptotic $(1-\alpha)$ confidence interval for $\mathrm{PW}_{p,\lambda}(r,s)$.

We apply the proposed interval estimation method to handwritten letter images from the Modified National Institute of Standards and Technology (MNIST) data set (`http://yann.lecun.com/exdb/mnist/`). The data set contains images of 576 pixels for handwritten digits from zero to nine. Because the distributions generating the images of each digit are likely to have low-dimensional structures, the PRW distance is expected to capture the differences between them effectively. Based on the above result, we construct 0.95 confidence intervals for the regularized PRW distances between pairs of digits. Specifically, we use $n = m = 892$ images of the digits zero, one, four, seven ,and nine, and extract 128-dimensional features of each image using a convolution neural network (CCN), as outlined in Lin et al. (2020). Then, we estimate the global intrinsic dimension of the feature data using the maxLikLocalDimEst function in the R package *intrinsicDimension* Johnsson (2019), obtaining an estimate of 6.77. Based on this estimate, we set the projection dimension to 7 and the order to $p = 2$. We then construct the 0.95 confidence intervals using $B = 1{,}000$ bootstraps with replacement $n^{4/5} \approx 230$. The regularized PRW distance is calculated using the Riemannian optimization method proposed by Lin et al. (2020).

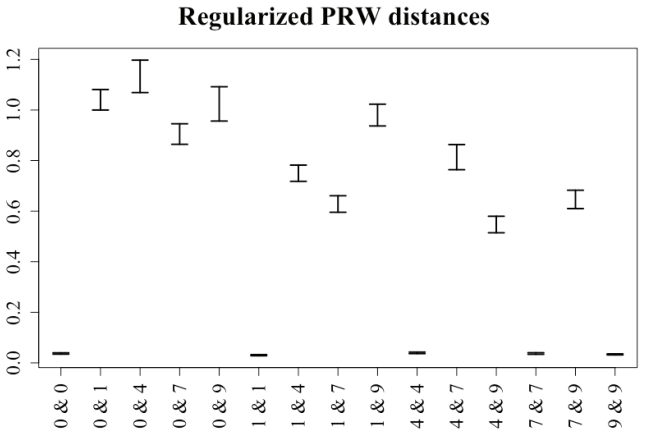

Figure 2. The 0.95 confidence intervals for the regularized PRW distance between handwritten digits. Intervals for the same digits are calculated by splitting the data set into two groups. Intervals are normalized by setting the lower bound for zero and one to one.

Figure 2 shows the results. The distances between digits one and seven and between digits four and nine are smaller than those between digits zero and one and between zero and four. Moreover, the distances between the same digits are quite small. These results are consistent with our intuition.

Furthermore, we add Gaussian noise with a standard deviation of $\sigma = 1, 5, 10$ to the feature data, and again construct 0.95 confidence intervals for the regularized PRW distances. For comparison, we also construct 0.95 confidence intervals for the original Wasserstein distances (Sommerfeld and Munk (2018)). The results are shown in Figure 3. The interval estimates of the regularized PRW distance are less affected by the increase in the variance of the noise than are those of the Wasserstein distance. This result implies that the PRW distance is more robust to noise than the original Wasserstein distance is when the data set has a low-dimensional structure.

6. Conclusion

This study investigates statistical inference for the IPRW and regularized PRW distances. Although these projection-based Wasserstein distances are practical for many machine learning tasks, their inferential tools are not well established. We derive the limit distributions of the empirical versions of these distances on finite spaces by showing their directional Hadamard differentiability. We also show that, although the naive bootstrap fails for these distances, the rescaled bootstrap is consistent.

There are several promising directions for future research. First, our theoretical results are limited to finitely supported measures, and it would be

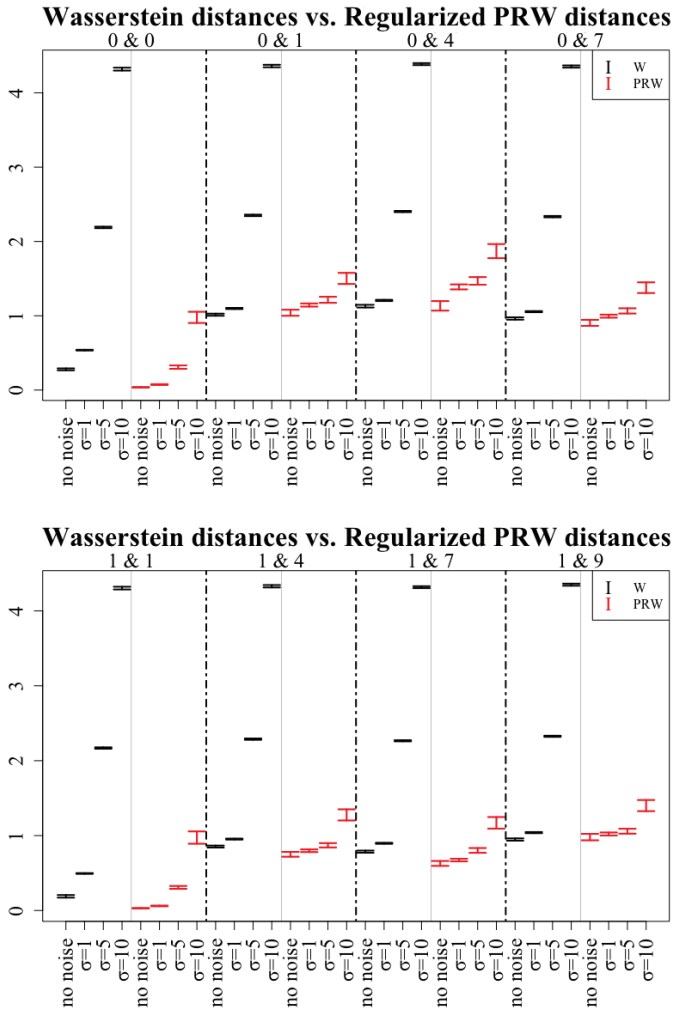

Figure 3. The 0.95 confidence intervals for the regularized PRW and Wasserstein distance between handwritten digits with Gaussian noise. The intervals for the same digits are calculated by splitting the data set into two groups. For each distance, the intervals are normalized by setting the lower bound for zero and one to one.

worthwhile extending them to more general settings. Second, an appropriate choice of the replacement number of the rescaled bootstrap or projection dimension of the PRW distance is important in practice. Developing data-driven methods to choose these values is left to future research.

Supplementary Material

The online Supplementary Material contains the proofs of theorems, propositions, and lemmas presented in the main paper, as well as additional simulation results.

Acknowledgments

The authors are grateful for the very constructive comments of the two anonymous referees.

References

Bigot, J., Cazelles, E. and Papadakis, N. (2019). Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics* **13**, 5120–5150.

Bonneel, N., Rabin, J., Peyré, G. and Pfister, H. (2015). Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* **51**, 22–45.

Carriere, M., Cuturi, M. and Oudot, S. (2017). Sliced Wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning* **PMLR 70**, 664–673.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems* **26**, 2292–2300.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. and Rodríguez-Rodríguez, J. M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics* **27**, 1230–1239.

Del Barrio, E., Giné, E. and Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11**, 131–189.

del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability* **47**, 926–951.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S. et al. (2019). Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10640–10648.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields* **95**, 125–140.

Fiacco, A. V. (1983). *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic Press.

Freitag, G. and Munk, A. (2005). On Hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships. *Journal of Multivariate Analysis* **94**, 123–158.

Goldfeld, Z. and Greenewald, K. (2020). Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistic* **PMLR 108**, 3327–3337.

Johnsson, K. (2019). *intrinsicDimension: Intrinsic Dimension Estimation*. R package version 1.2.0.

Klatt, M., Tameling, C. and Munk, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science* **2**, 419–443.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R. and Rohde, G. K. (2019). Generalized sliced Wasserstein distances. *arXiv:1902.00434*.

Kolouri, S., Pope, P. E., Martin, C. E. and Rohde, G. K. (2019). Sliced Wasserstein Auto-Encoders. In *International Conference on Learning Representations (ICLR 2019)*, 1–19. Conference Paper.

Kolouri, S., Zou, Y. and Rohde, G. K. (2016). Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5258–5267.

Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K. (2015). From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning* **PMLR 37**, 957–966.

Le, T., Yamada, M., Fukumizu, K. and Cuturi, M. (2019). Tree-sliced variants of Wasserstein distances. *arXiv:1902.00342*.

Lin, T., Fan, C., Ho, N., Cuturi, M. and Jordan, M. I. (2020). Projection robust Wasserstein distance and Riemannian optimization. *arXiv:2006.07458*.

Lin, T., Zheng, Z., Chen, E. Y., Cuturi, M. and Jordan, M. I. (2021). On projection robust optimal transport: Sample complexity and model misspecification. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics* **130**, 262–270.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A. and Stöter, F.-R. (2019). Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning* **97**, 4104–4113.

Manole, T., Balakrishnan, S. and Wasserman, L. (2019). Minimax confidence intervals for the sliced Wasserstein distance. *arXiv:1909.07862*.

Mena, G. and Weed, J. (2019). Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. *arXiv:1905.11882*.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 223–241.

Nguyen, K., Ho, N., Pham, T. and Bui, H. (2020). Distributional sliced-Wasserstein and applications to generative modeling. *arXiv:2002.07367*.

Niles-Weed, J. and Rigollet, P. (2019). Estimation of Wasserstein distances in the spiked transport model. *arXiv:1909.07513*.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application* **6**, 405–431.

Paty, F.-P. and Cuturi, M. (2019). Subspace robust Wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning* **PMLR 97**, 5072–5081.

Peyré, G. and Cuturi, M. (2019). *Computational Optimal Transport: With Applications to Data Science*, 355–607. Now Publisher, Boston.

Rabin, J., Peyré, G., Delon, J. and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 435–446. Springer.

Ramdas, A., Trillos, N. G. and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19**, 47.

Römisch, W. (2004). Delta method, infinite dimensional. *Encyclopedia of Statistical Sciences* **3**. Web: `https://doi.org/10.1002/0471667196.ess3139`.

Rubner, Y., Tomasi, C. and Guibas, L. J. (2000). The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision* **40**, 99–121.

Sandler, R. and Lindenbaum, M. (2011). Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 1590–1602.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A. et al. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)* **34**, 1–11.

Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 219–238.

Tameling, C., Sommerfeld, M. and Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability* **29**, 2744–2781.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Villani, C. (2009). *Optimal Transport: Old and New*. Springer.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25**, 2620–2648.

Zhang, M., Liu, Y., Luan, H., Sun, M., Izuha, T. and Hao, J. (2016). Building Earth Mover's Distance on bilingual word embeddings for machine translation. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2870–2876.

Ryo Okano

KIS, The Unviersity of Tokyo, Meguro-ku, Tokyo 153-0041, Japan.

E-mail: okano-ryo1134@g.ecc.u-tokyo.ac.jp

Masaaki Imaizumi

KIS, The Unviersity of Tokyo, Meguro-ku, Tokyo 153-0041, Japan.

E-mail: imaizumi@g.ecc.u-tokyo.ac.jp

(Received February 2022; accepted July 2022)